Winning Space Race with Data Science

Mayuko Kondo
Data Scientist, SpaceY

October 27, 2021

Outline

1. Executive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion
6. Appendix

1. Executive Summary

We have done:

- SpaceX flight data collected by API and web scraping, and performed EDA using visualization and SQL
- Made interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

We have found that:

- The flight without gridfins is more likely to fail
- After controlling all major factors to determine outcome, whether reusing first stage or not does not affect the landing outcome
- Reusing first stage is the good strategy to take for reducing cost of rocket launch in terms of statistical analysis of landing outcome

We think that:

- Whether SpaceY should reuse the first stage or not depends also on:
  - The actual cost of reusing the first stage; collection and maintenance cost vs producing the new
  - Durability; how many times we can reuse without affecting launch outcome etc.

2. Introduction

Background

- Rocket companies are making space travel affordable for everyone:
  - Virgin Galactic is providing suborbital spaceflights
  - Rocket Lab is a small satellite provider
  - Blue Origin manufactures sub-orbital and orbital reusable rockets
  - SpaceX sending spacecraft to ISS, Starlink to provide satellite internet, and manned missions
- Reusing the first stage is the key to reduce the cost of rocket launches

Question

Q1. What can determine the rocket launch outcomes; success or failure?

Q2. After controlling all other factors determining outcome, does the reuse of the first stage affect the outcome of the rocket launch?

Q3. Is reusing first stage worth?

Section 1

Methodology

3. Methodology

Summary

3.1. Data collection methodology:

- Collecting data by API and web scraping

3.2. Perform data wrangling

- Understanding and cleaning data for further analyses

3.3. Perform exploratory data analysis (EDA) using visualization and SQL

3.4. Perform interactive visual analytics using Folium and Plotly Dash

3.5. Perform predictive analysis using classification models

- Compare performance of models; Logistic Regression, Decision Tree, K Nearest Neighbor, Support Vector Machine
- Learn the determinants of successful outcome

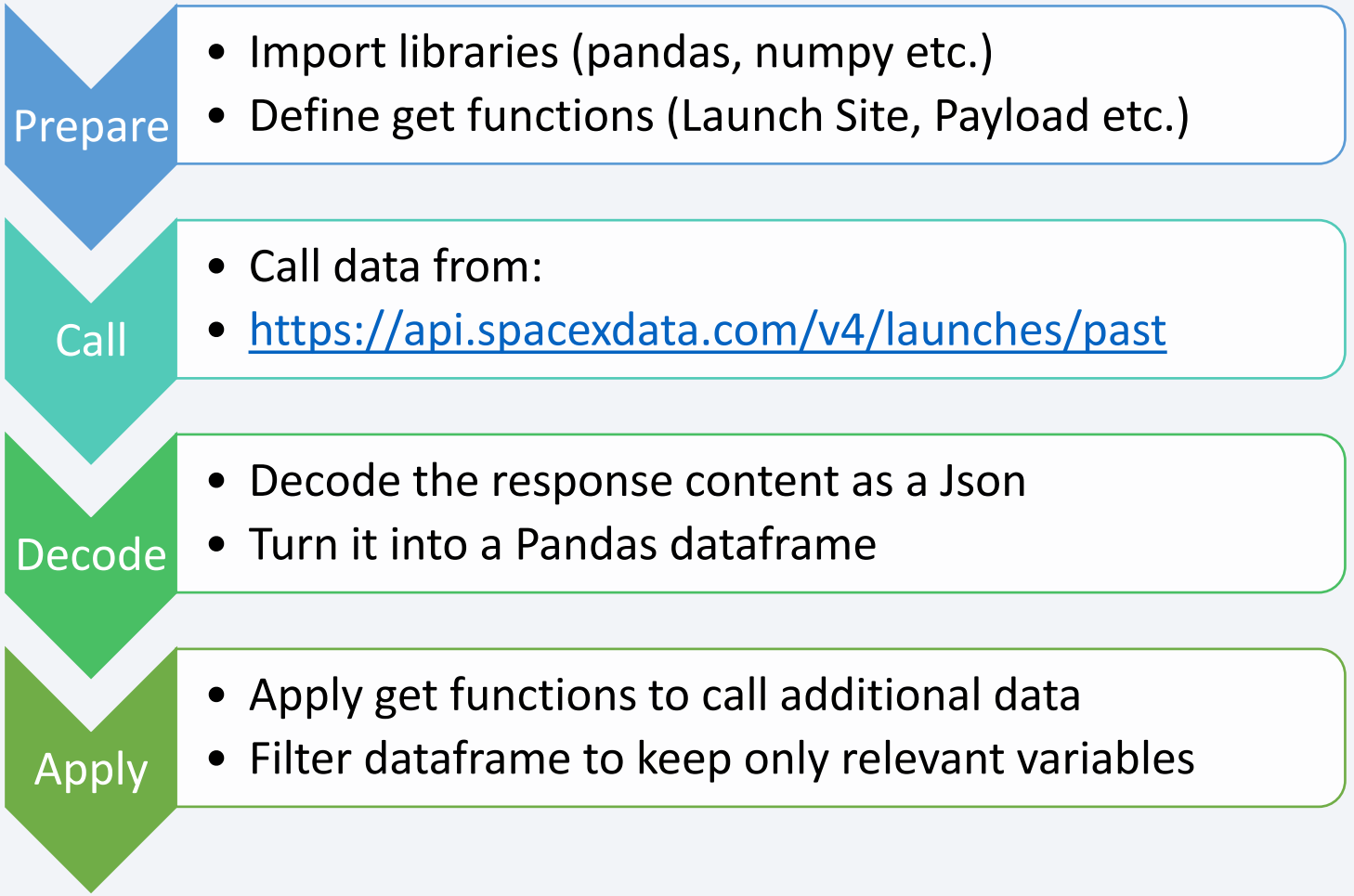
3.1. Data Collection

API	Web Scraping
• Direct access to data set • Extract data from required application • Limited access to all the available public data • Well-formatted data available • Took 4 steps; prepare, call, decode, and apply	• Indirect access to data set • Extract data from any website • Fewer limitation of access • Real-time data is available • Able to customize data • Anonymous • Took 4 steps; prepare, read, collect, and encode

➢ We collected data by both API and Web Scraping.

3.1.(i). Data Collection – SpaceX API

- We made a get request to the SpaceX API, collected data, and made sure the data is in the correct format from an API.
- GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/1.%20Data%20Collection%20API%20lab.ipynb



3.1.(ii). Data Collection – Web Scraping

- We performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
- GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/2.%20Data%20Collection%20with%20Web%20Scraping.ipynb

Prepare
- Import libraries (pandas, BeautifulSoup etc.)
- Define functions (Payload Mass, extract_column etc.)

Read
- Read HTML data from the web: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Collect
- Create a BeautifulSoup object from the HTML data
- Collect column names from the HTML table header

Encode
- Create dictionary with keys from extracted column names.
- Turn the dictionary into a Pandas dataframe

3.2. Data Wrangling

1. Eliminate irrelevant data
   - Keep only one data per one flight
   - Keep only Falcon9 data
2. Check data structure
   - Check data values
   - Check missing values
   - Check data type
3. Enquire key variables
   - Number of launches by site
   - Number of occurrence by orbit
4. Generate target variable
   - Generate outcome dummy variable, called ‘Class’, from landing outcome variable

- GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/3.%20EDA%20lab.ipynb

3.3.(i). EDA with Data Visualization

- Check: what explains the different success rate of each launch site
  - Show success rate by launch site
  - Scatter plot to show Flight Number vs. Launch Site
  - Scatter plot to show Payload vs. Launch Site
- Check: what explains the different success rate of each orbit type
  - bar chart for the success rate of each orbit type
  - Scatter plot to show flight number vs. Orbit type
  - Scatter plot to show payload vs. orbit type
- Check: yearly trend of average success rate

GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/5.%20EDA%20with%20Visualization%20tool.ipynb

3.3.(ii). EDA with SQL

- Enquiring the detail and relation of landing outcome, customer, payload mass, booster version, year, and landing type such as drone ship, ground pad etc.
- We performed EDA with SQL to know the detail of data relating to the landing outcome so that we understand the situations of the rocket flights well and would not miss the important condition or insight during analyses

GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/4.%20EDA%20with%20SQL.ipynb

3.4.(i). Build an Interactive Map with Folium

- We marked the 4 launch sites with circle markers on the world map
- We added markers showing red for the failed launch and green for the successful launch with the pop-up markers of total counts on the map of each site
- We added lines from each launch site to the nearest coastline, city, highway, and railway
- Then calculated the distance of each line, and added the distance of km to the map
- We did above to understand the geographical conditions of the launch sites, which potentially affect the launch outcome

GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/6.%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

3.4.(ii). Build a Dashboard with Plotly Dash

- Built the Pi-chart showing launch success count for all sites
- Built the Pi-chart showing launch success ratio (success/failure) by site with dropdown menu showing the launch sites
- Built the scatter plot showing the relation between payload mass and launch outcome with adjustable payload mass range selection bar
- We explored above to know whether the launch site or payload mass determine the launch outcome

GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/7.%20Dashboard_Application_Plotly_Dash.py

3.5. Predictive Analysis (Classification)

Logistic Regression	SVM	Decision Tree	KNN
Optimize model using train and test data	Optimize model using train and test data	Optimize model using train and test data	Optimize model using train and test data
Compute accuracy of the model	Compute accuracy of the model	Compute accuracy of the model	Compute accuracy of the model
Examine confusion matrix	Examine confusion matrix	Examine confusion matrix	Examine confusion matrix

- Compare the accuracy of each model to apply the best optimized model to predict launch outcome
- Then enquire the determinants of launch outcome using the best model

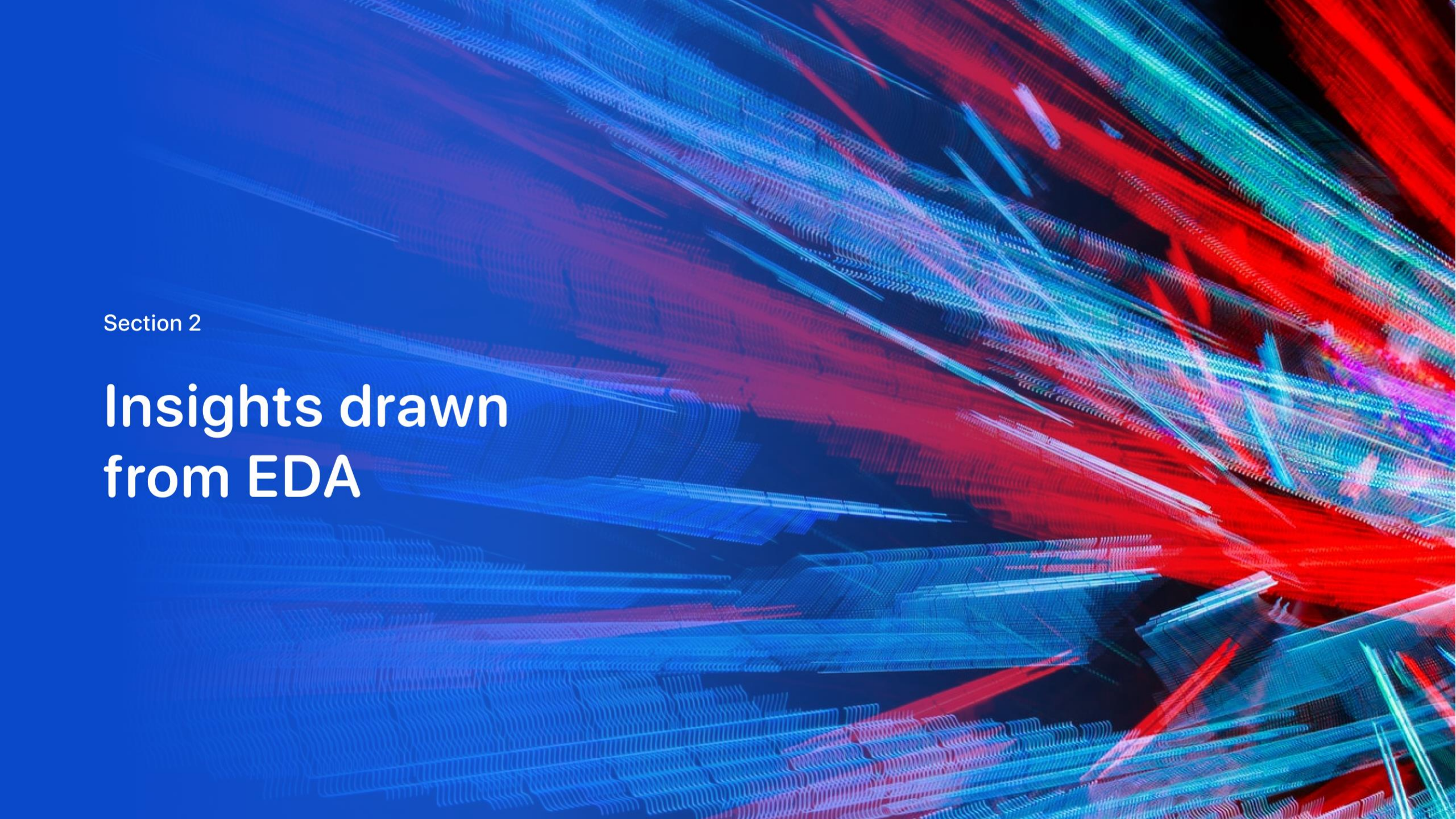
GitHub URL: https://github.com/mayuko-kondo/spacex_data_science/blob/master/8.%20Machine%20Learning%20Prediction.ipynb

4. Results

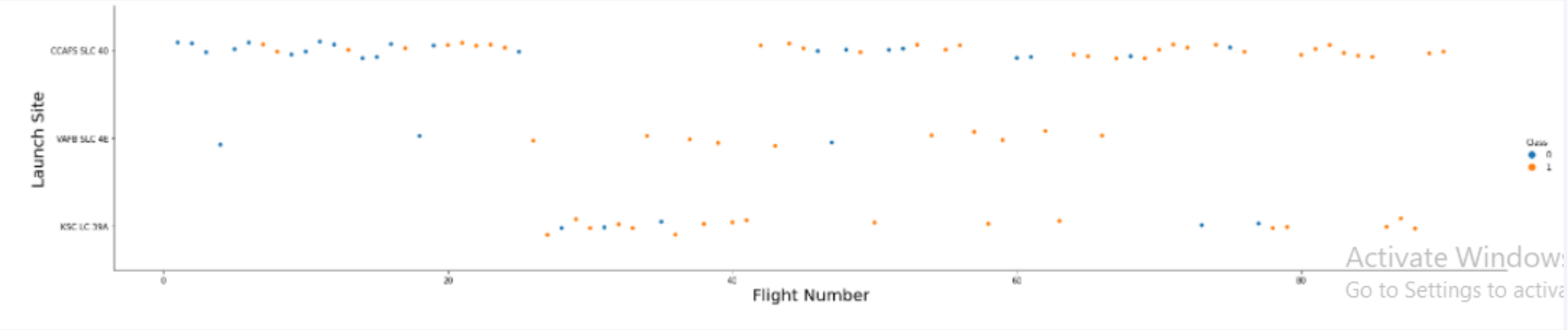
Summary

Section 2

Insights drawn from EDA

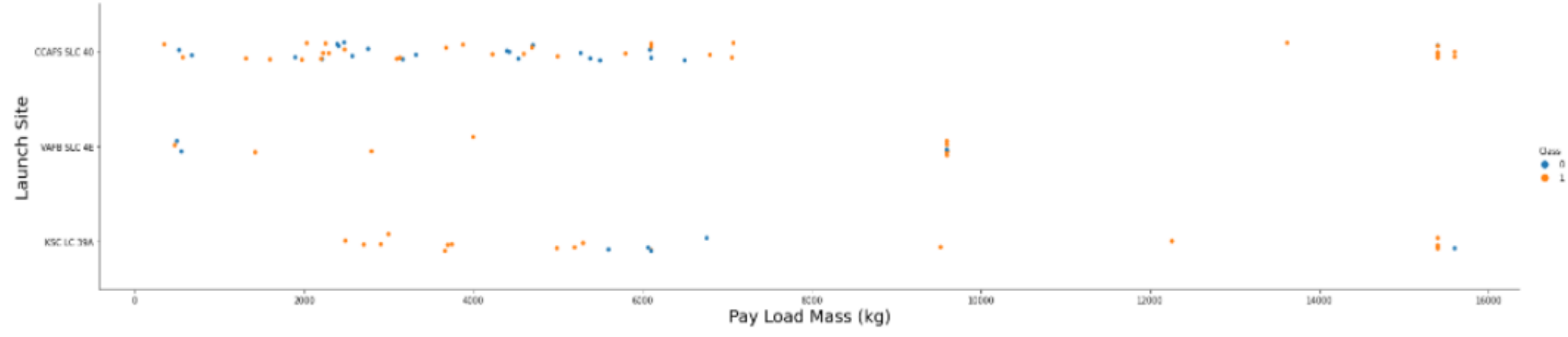


4.1.(i). Flight Number vs. Launch Site



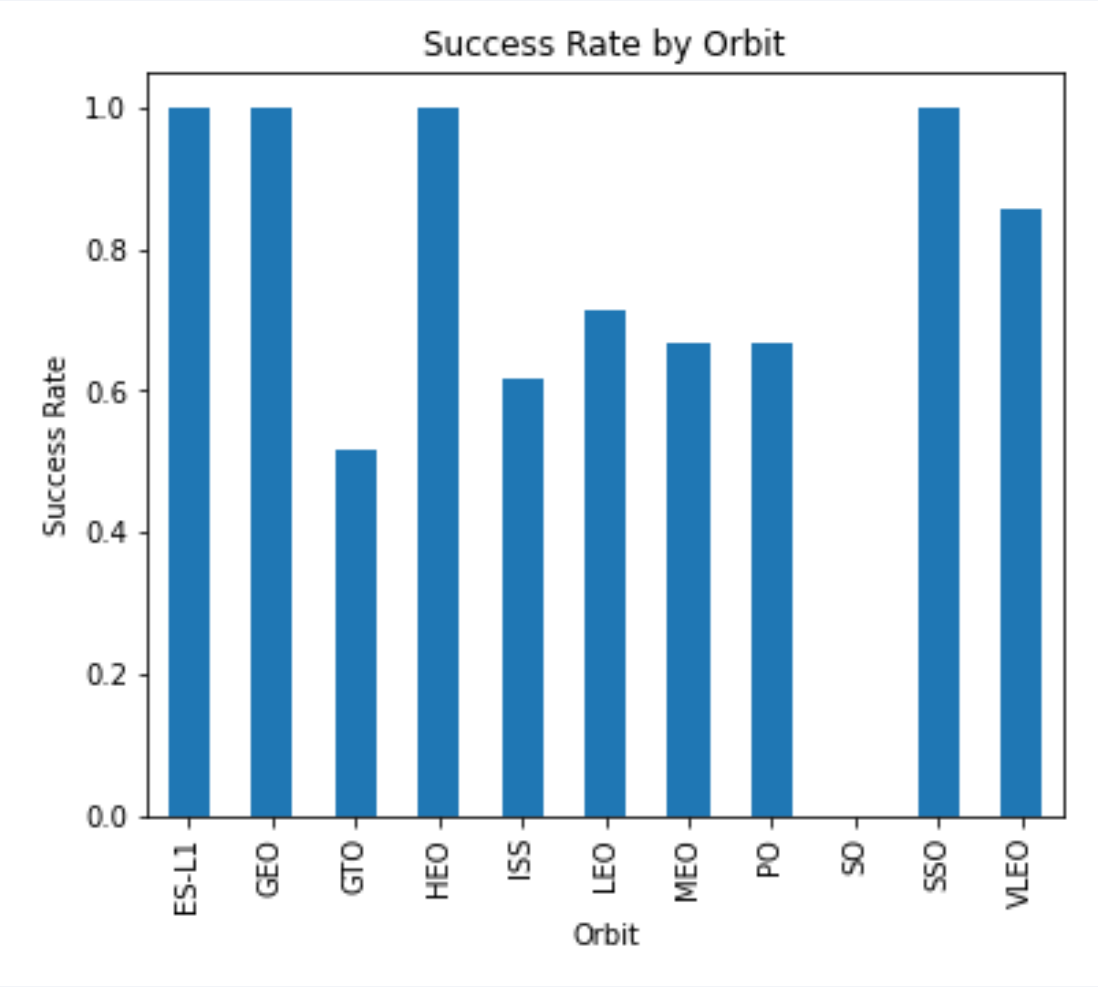
- X-axis showing flight number, Y-axis showing launch site, CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A, from top to bottom.
- Success rate increases as flight number increases in all sites
- CCAFS SLC 40 has the largest number of launches
- Since spaceX started rocket launch at CCAFS SLC 40, it has a greater number of failures before it started having good success rate than other two sites.

4.1.(ii). Payload vs. Launch Site



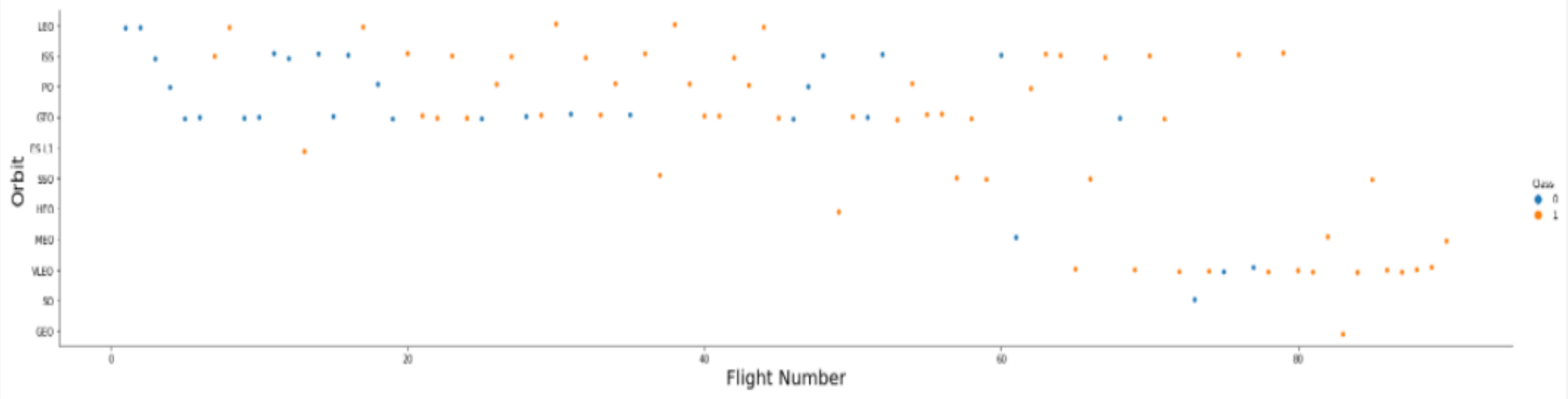
- X-axis showing payload mass (kg), Y-axis showing launch site, CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A, from top to bottom.
- VAFB SLC 4E site has no rockets launched for heavy payload mass greater than 10000kg.
- CCAFS SLC 40 has greater success rate with heavy payload mass greater than 7000kg.
- KSC LC 39A has greater success rate with lower and heavier payload mass; smaller than 5000kg and greater than 8000kg.

4.1.(iii). Success Rate vs. Orbit Type



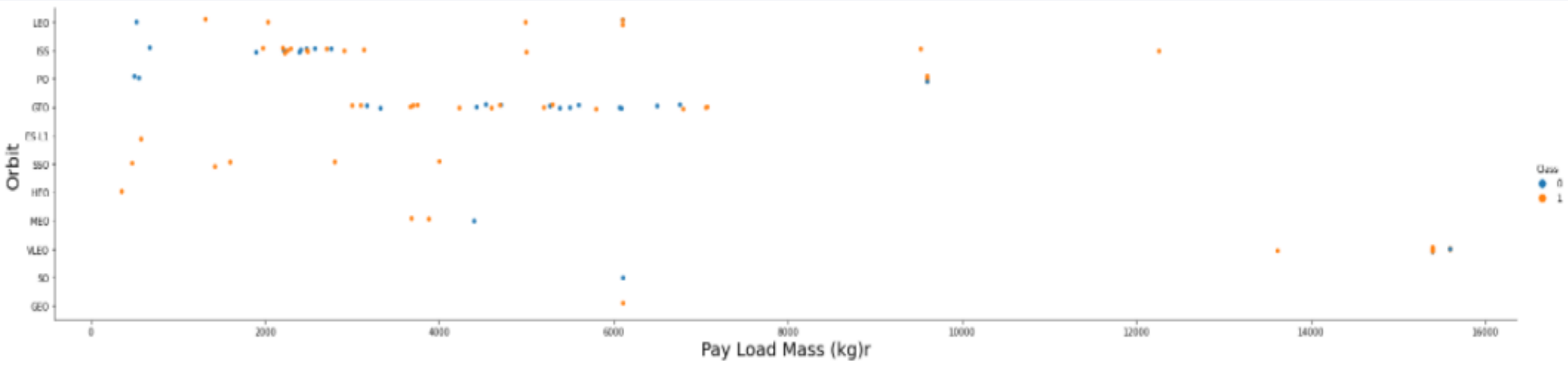
- X-axis showing each orbit type, Y-axis showing success rate.
- ES-L1, GEO, HEO, and SSO have 100% success rate followed by VLEO with over 80%.
- GTO has the lowest success rate around 50%.
- ISS, LEO, MEO, and PO all have success rate of around 60-70%.
- Orbit type has more variation of success rate than launch site.

4.1.(iv). Flight Number vs. Orbit Type



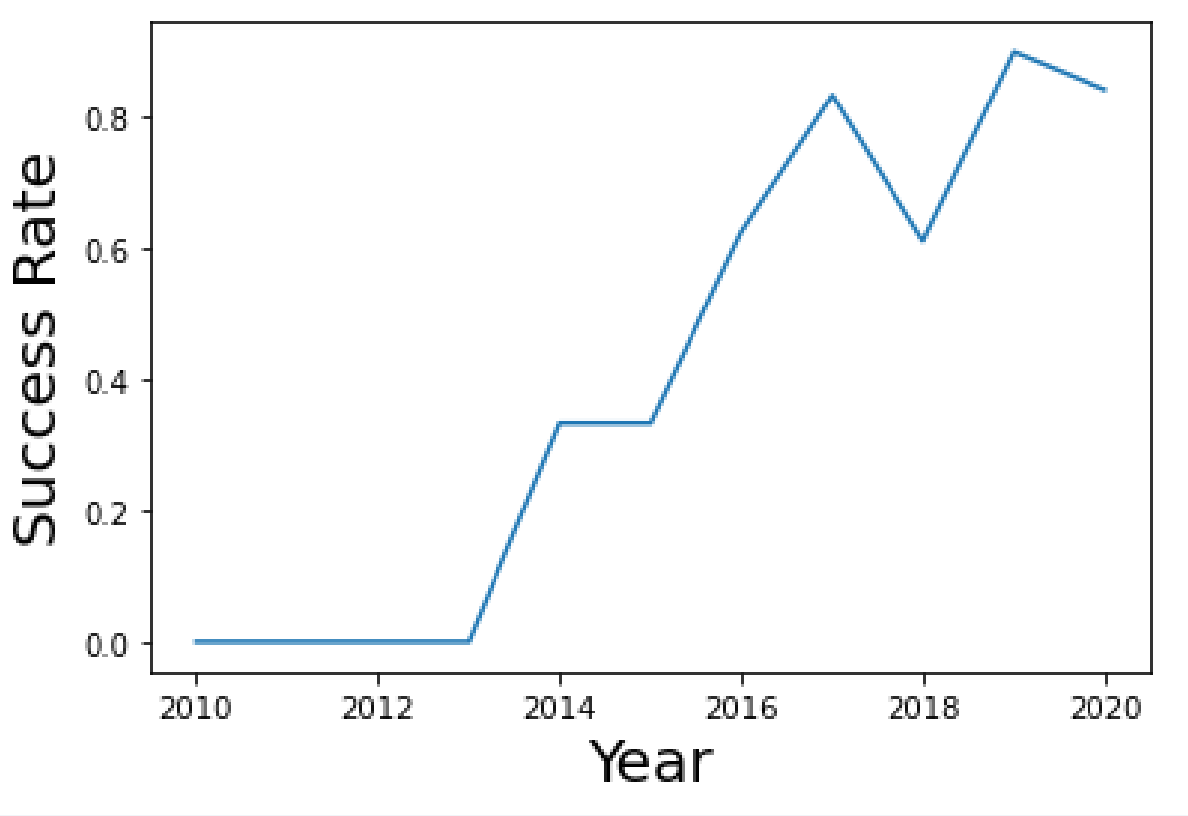
- X-axis showing flight number, Y-axis showing orbit type.
- LEO orbit shows the Success appears as the number of flight increases
- GTO orbit shows no relationship between successful launch and flight number
- For most orbits, success rate increases as flight number increases.

4.1.(v). Payload vs. Orbit Type



- X-axis showing payload mass (kg), Y-axis showing orbit type.
- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS orbits.
- GTO orbit shows no clear distinction among payload mass rate because it shows both positive landing rate and negative landing regardless of the payload mass.

4.1.(vi). Launch Success Yearly Trend



- X-axis showing year, Y-axis showing success rate.
- Success rate continuously increased since 2013 until 2020.

4.1.(vii). All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-858
Done.
```

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- There are four distinct launch sites; CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

4.1.(viii). Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' limit 5
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	Landing _Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- There are five records with has the name of the launch site starting with CCA in SPACEXTBL data set.

4.1.(ix). Total Payload Mass

```
%sql SELECT SUM(payload_mass__kg_) AS sum_payload_kg_NASACRS FROM SPACEXTBL WHERE CUSTOMER LIKE '%NASA (CRS)%'
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3
Done.
```

sum_payload_kg_nasacrs
48213

- The sum of payload mass carried by boosters launched by NASA(CRS) is 48213 kg.

4.1.(x). Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) AS average_payload_kg_F9v1_1 FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1'
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505
Done.
```

average_payload_kg_f9v1_1
2928

- The average payload mass carried by booster version F9 v1.1 is 2928 kg.

4.1.(xi). First Successful Ground Landing Date

```
%sql SELECT * FROM SPACEXTBL WHERE DATE = (SELECT min(DATE) FROM SPACEXTBL WHERE "LandingOutcome" LIKE 'Success (ground pad)')
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	LandingOutcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- The date of first successful landing outcome in ground pad is December 22, 2015.

4.1.(xii). Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT booster_version FROM SPACEXTBL WHERE "LandingOutcome" LIKE 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND p
ayload_mass__kg_ < 6000
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- There are 4 booster versions which have successful landing in drone ship with payload mass between 4000 kg and 6000 kg; F9 FT B1021.2, F9 FT B1031.2, F9 FT B1022, and F9 FT B1026.

4.1.(xiii) Total Number of Successful and Failure Mission Outcomes

```
#       mission_outcome LIKE 'Failure%'
%sql SELECT MISSION_SUCCESS_FAILURE, COUNT(*) FROM SPACEXTBL GROUP BY MISSION_SUCCESS_FAILURE
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.
Done.
```

mission_success_failure	2
FAILURE	1
SUCCESS	100

- There are one failure mission outcome and 100 success mission outcome.

4.1.(xiv). Boosters Carried Maximum Payload

```
%sql SELECT booster_version FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- There are 12 booster versions which have carried the maximum payload mass.

4.1.(xv). 2015 Launch Records

```
%sql SELECT DATE booster_version, launch_site, "LandingOutcome" FROM SPACEXTBL WHERE "LandingOutcome" LIKE 'Failure (drone ship)' AND yea
r(DATE) = 2015
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

booster_version	launch_site	LandingOutcome
2015-01-10	CCAFS LC-40	Failure (drone ship)
2015-04-14	CCAFS LC-40	Failure (drone ship)

- There are 2 records which failed in landing in drone ship in 2015.
- Both have the launch site, CCAFS LC-40.

4.1.(xvi). Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
%sql SELECT "LandingOutcome", COUNT(*) FROM SPACEXTBL WHERE DATE >= '2010-06-04' and Date <= '2017-03-20' GROUP BY "LandingOutcome" ORDER BY COUNT(*) DESC
```

```
 * ibm_db_sa://jjh06837:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

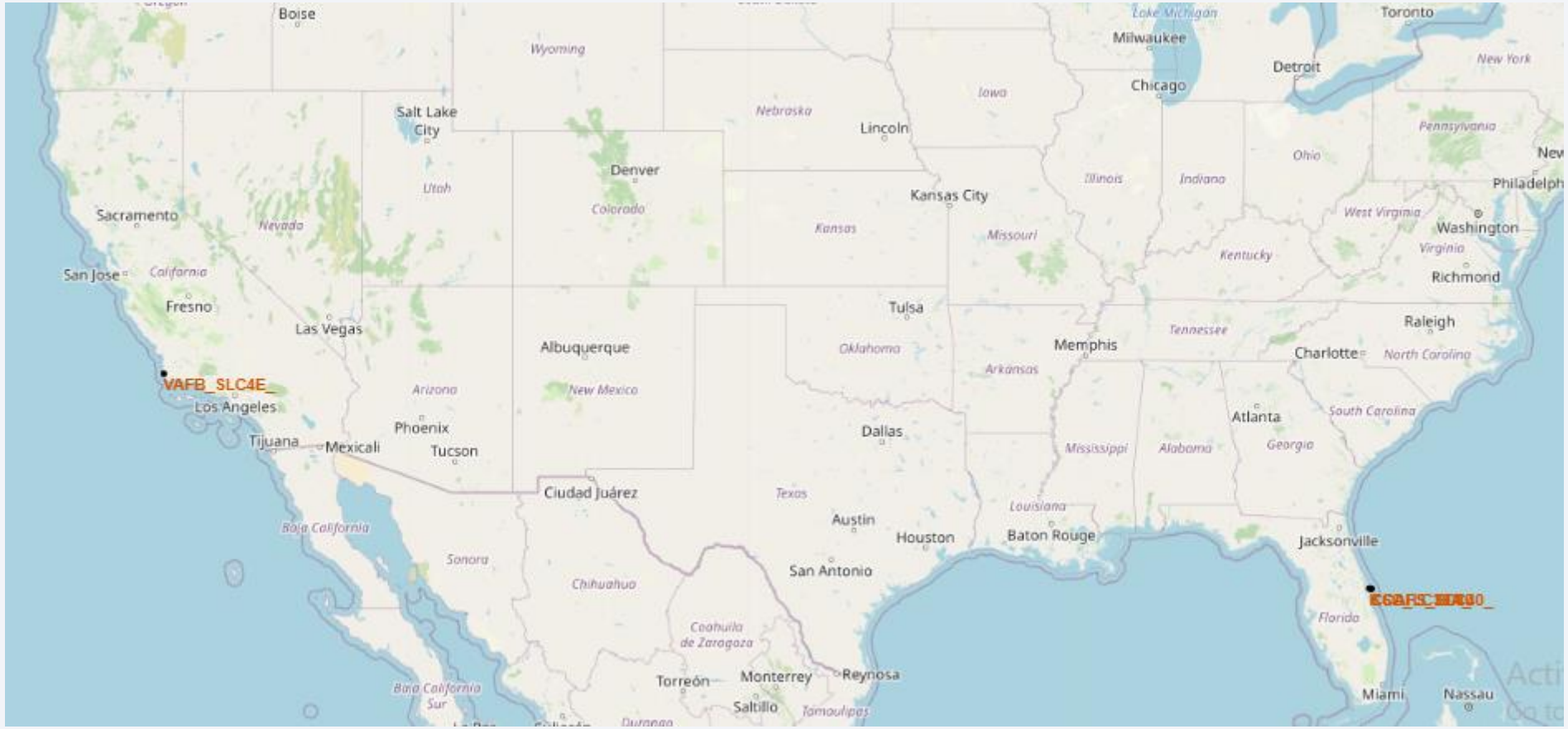
LandingOutcome	2
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- The highest number of landing outcome between 2010-06-04 and 2017-03-20 is no attempt of 10, followed by failure in drone ship of 5 and success in drone ship of 5.

Section 4

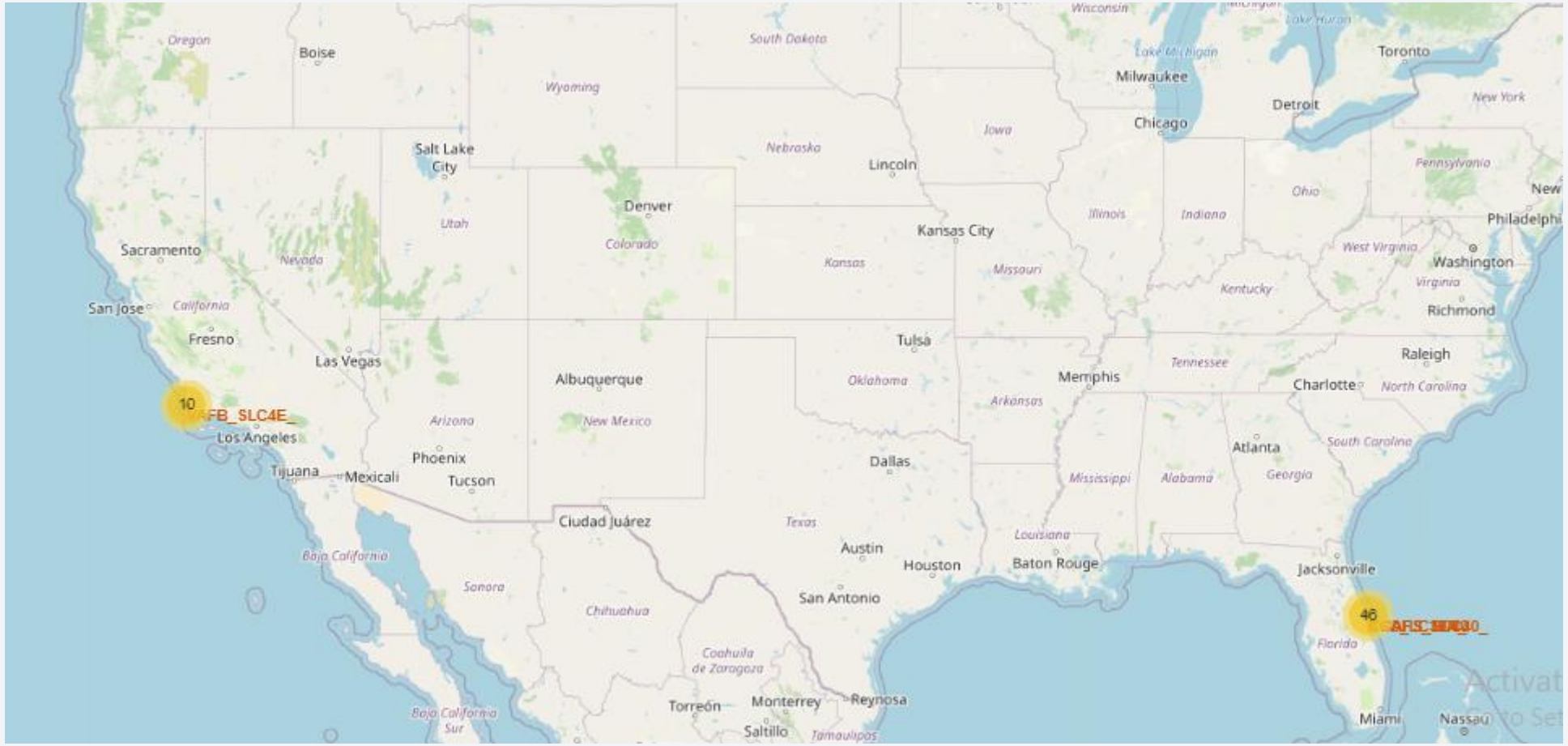
Launch Sites Proximities Analysis

4.1.(xvii). Launch Sites' Locations on Map



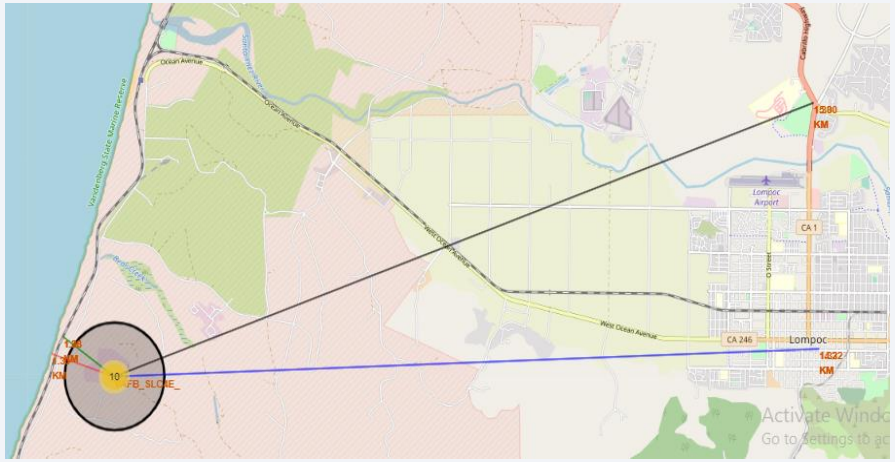
- Explain the important elements and findings on the screenshot

4.1.(xviii). Launch Sites and Outcomes on Map

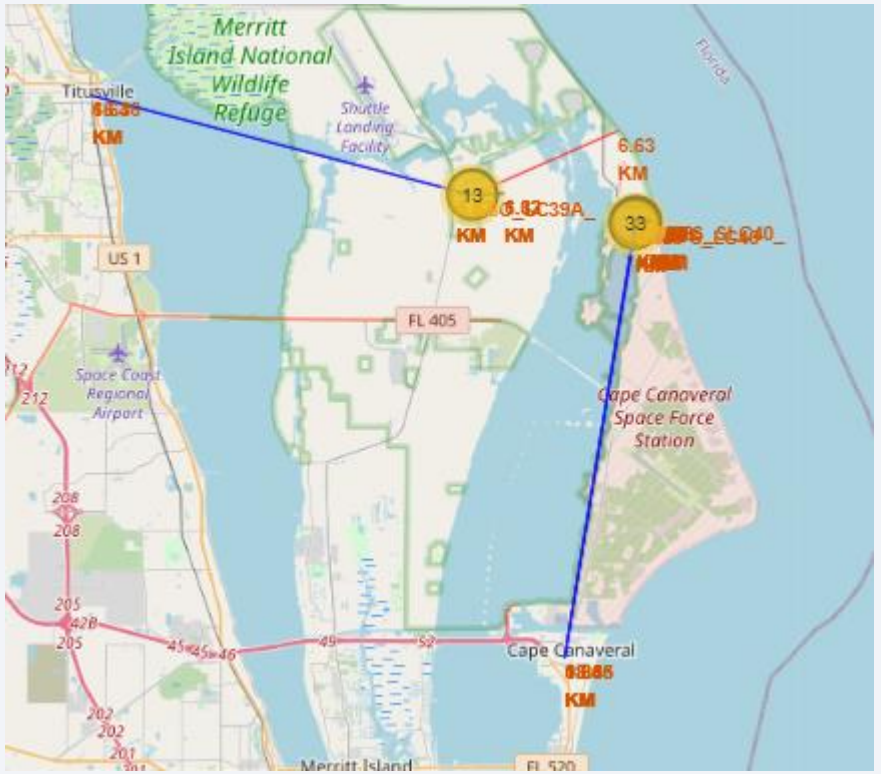


- Explore the folium map and make a proper screenshot to show the color-labeled launc
- Explain the important elements and findings on the screenshot

4.1.(xix). Location of Launch Sites by Distance



VAFB SLC-4E
Distance to coast: 1.33 km
Distance to city: 14.22 km
Distance to highway: 15.00 km
Distance to railway: 1.26 km



CCAFS LC-40
Distance to coast: 0.94 km
Distance to city: 18.45 km
Distance to highway: 0.65 km
Distance to railway: 0.00 km

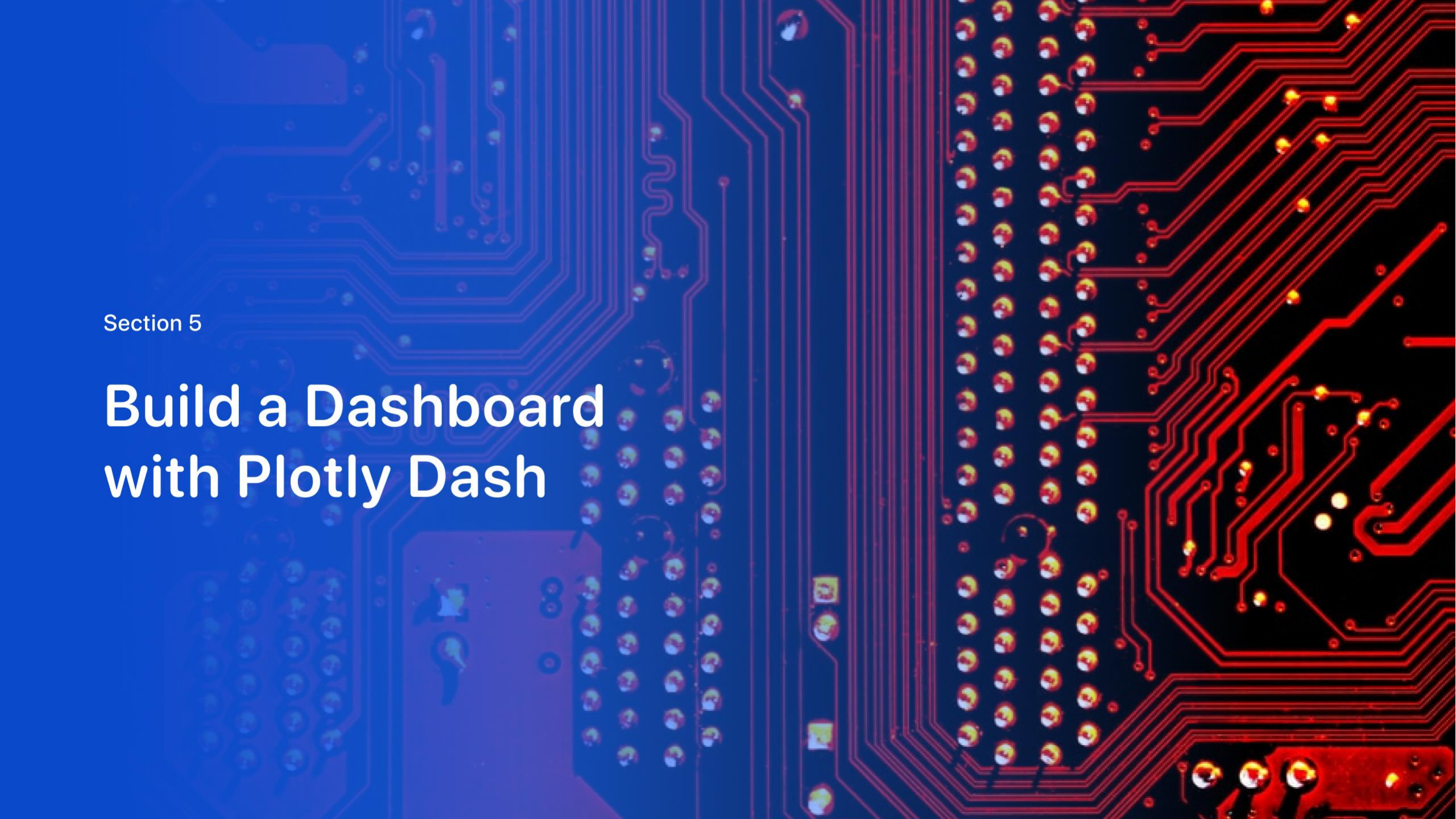
CCAFS SLC-40
Distance to coast: 0.86 km
Distance to city: 18.56 km
Distance to highway: 0.58 km
Distance to railway: 0.00 km

KSC LC-39A
Distance to coast: 6.63 km
Distance to city: 16.45 km
Distance to highway: 1.32 km
Distance to railway: 0.72 km

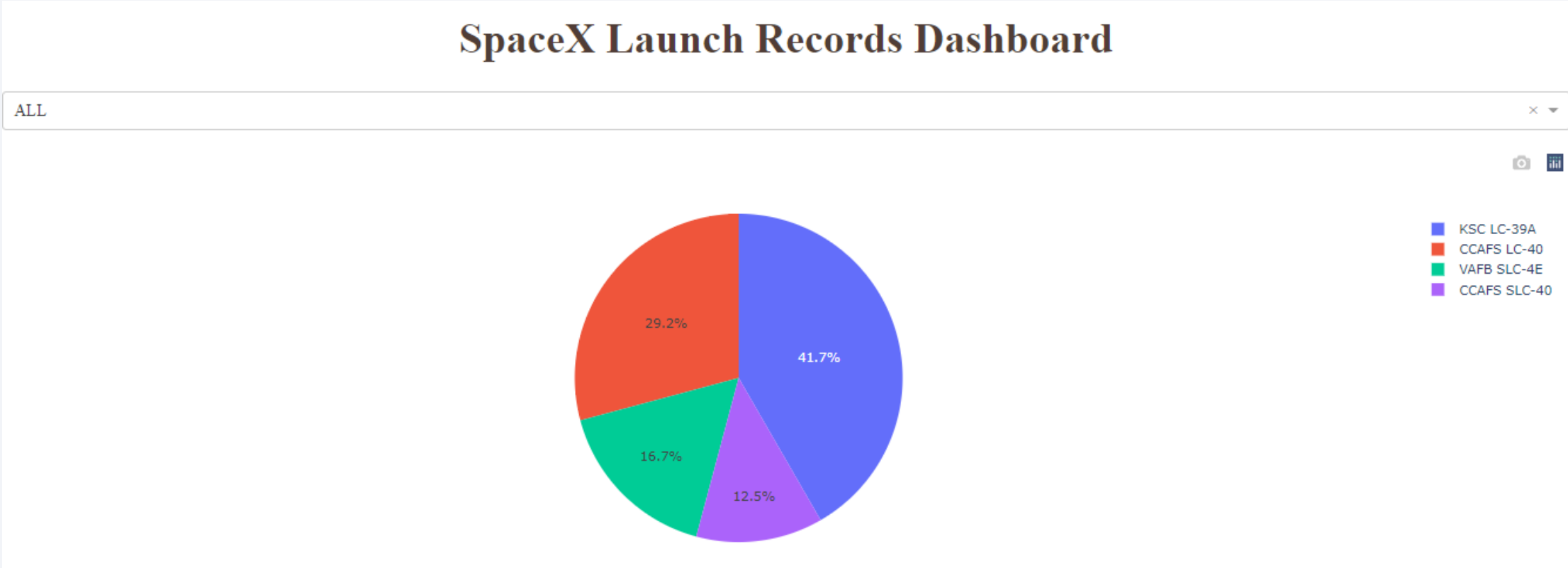
- Explore the generated folium map and show the screenshot of a selected launch site to it
- Explain the important elements and findings on the screenshot

Section 5

Build a Dashboard with Plotly Dash

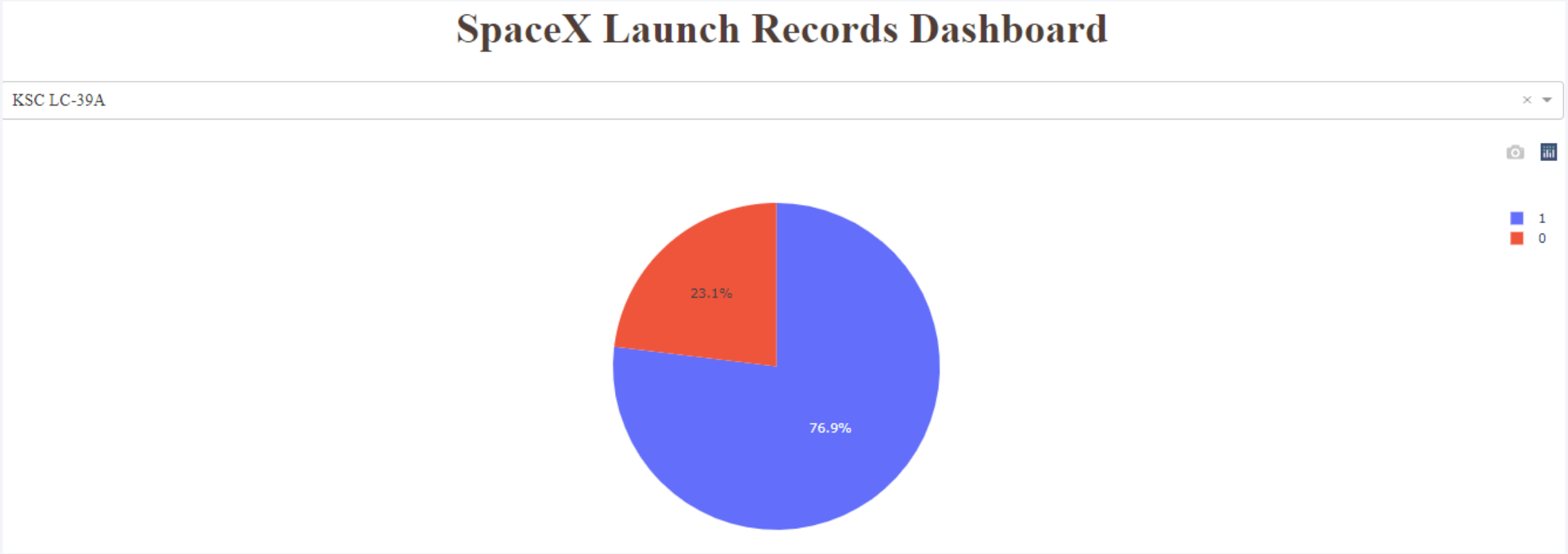


4.2.(i). Launch Success Count for All Sites



- KSC LC-39A has the largest successful launches. It accounts for 41.7% of all successful launches.
- CCAFS LC-40 has the second largest successful launches, which accounts for 29.2%.

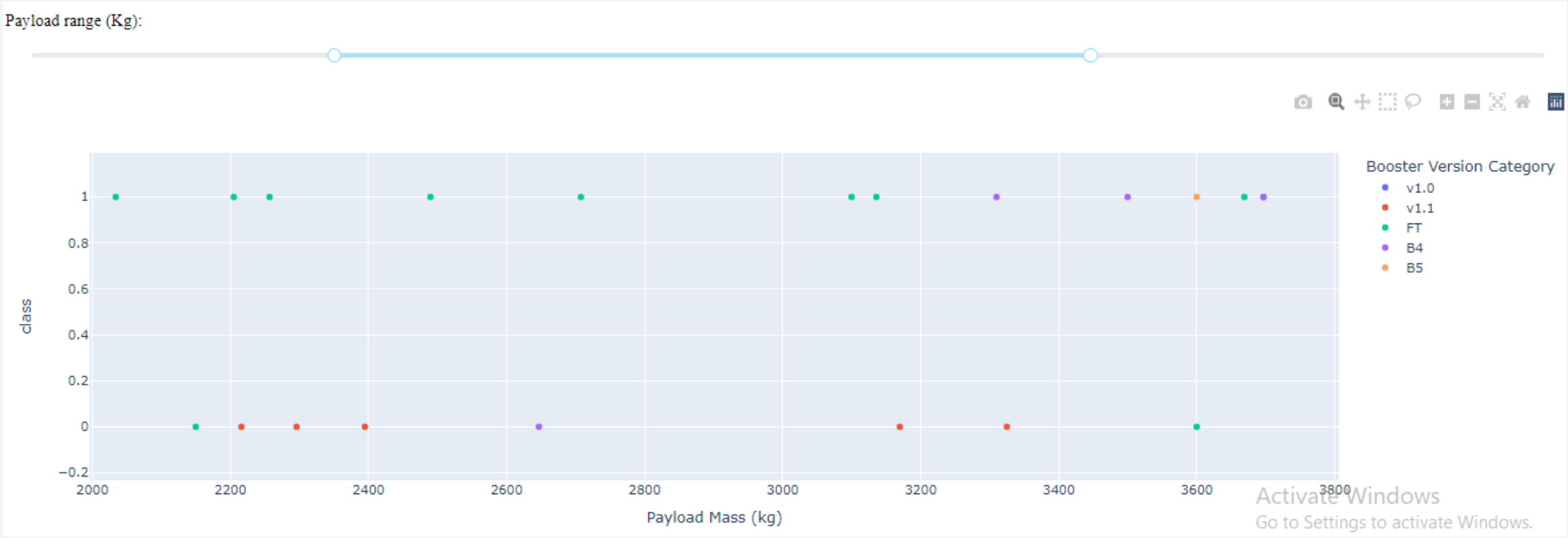
4.2.(ii). Launch Success Ratio by Site



- KSC LC-39A has the highest launch success ratio among all sites.
- The success ratio presents 76.9%.

4.2.(iii). Payload vs. Launch Outcome 1

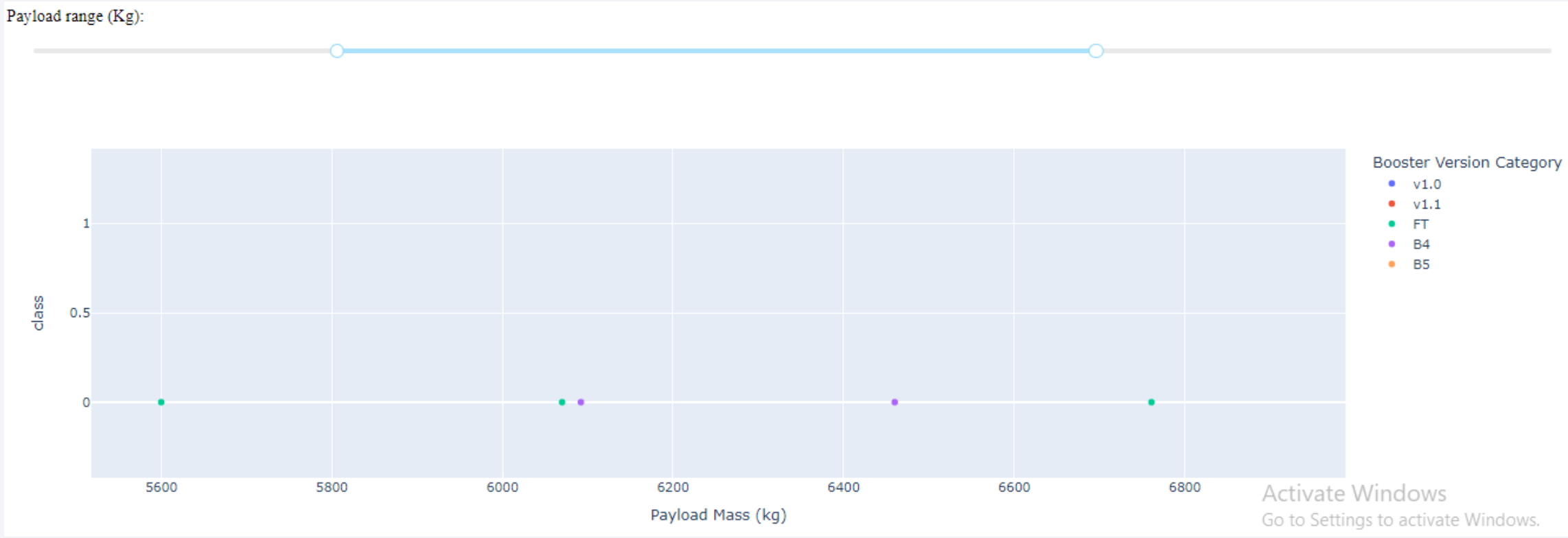
Payload Ranges with High Launch Success Rate



- Payload mass range 2000-3500kg has higher launch success rate than lower or higher payload mass ranges.

4.2.(iv). Payload vs. Launch Outcome 2

Payload Ranges with Low Launch Success Rate

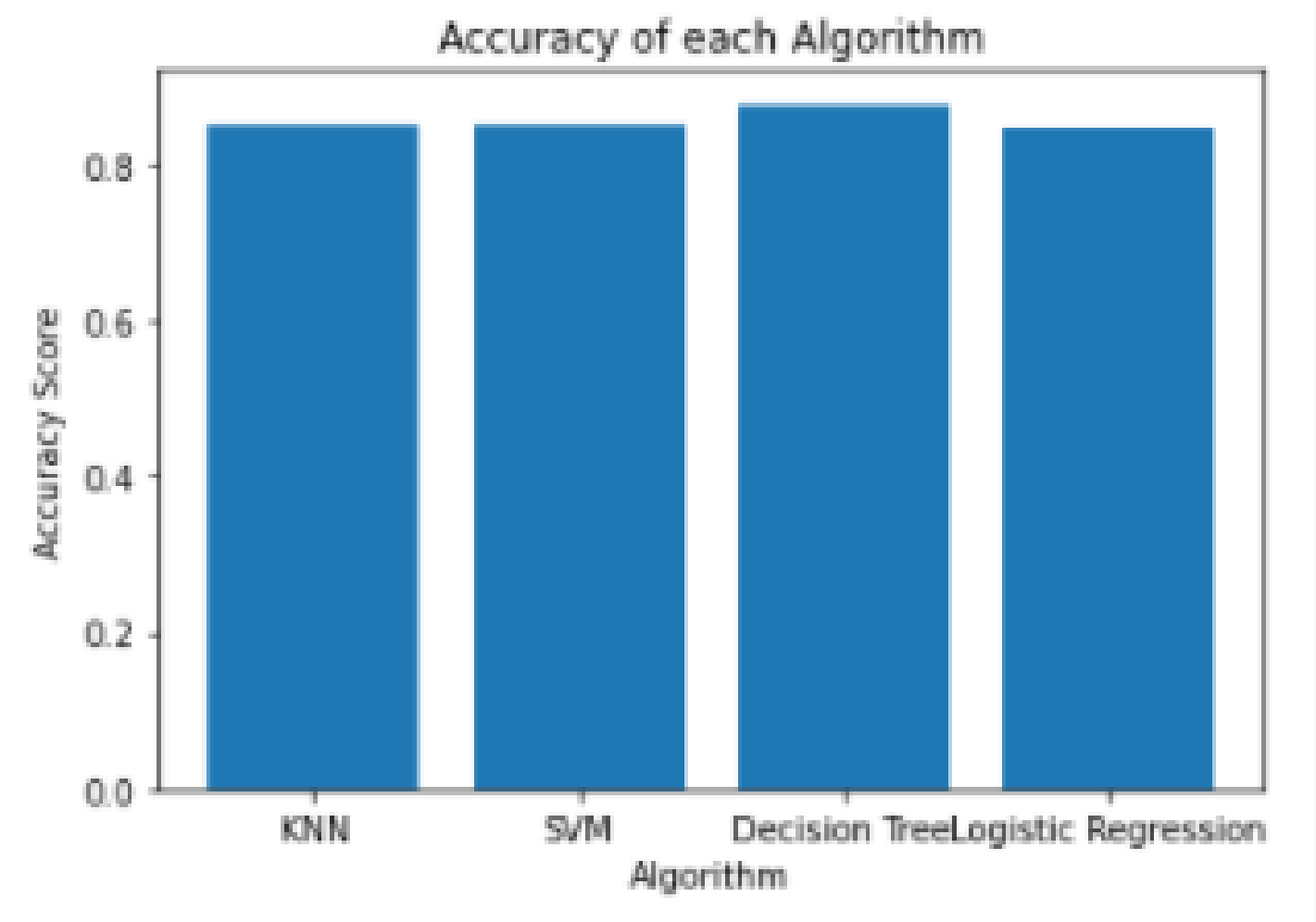


- Payload mass range 5600-6800kg has the lowest launch success rate.
- The range does not have a successful launch.

Section 6

Predictive Analysis (Classification)

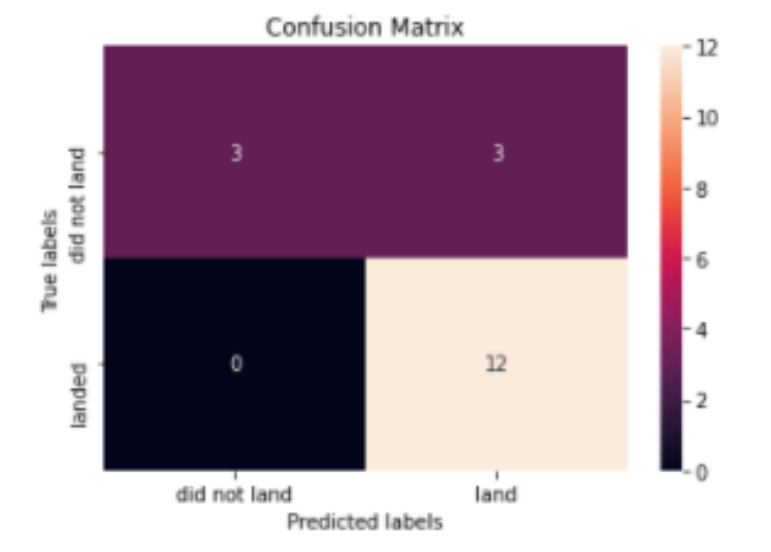
4.3.(i). Classification Accuracy



- Accuracy of built model:
  - KNN: 0.848
  - SVM: 0.848
  - Decision Tree: 0.877
  - Logistic Regression: 0.846
- Best algorithm is Decision Tree.
- Best set of parameters is:

  criterion: entropy
  max_depth: 4
  max_features: auto
  min_samples_leaf: 1
  min_samples_split: 5
  splitter: random

4.3.(ii). Confusion Matrix



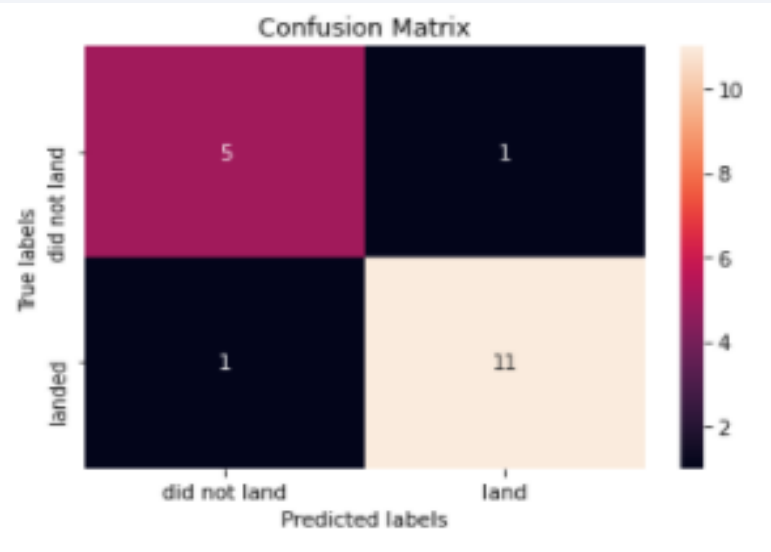
Logistic Regression

Counts of prediction:

True: 15

False Positive: 3

False Negative: 0



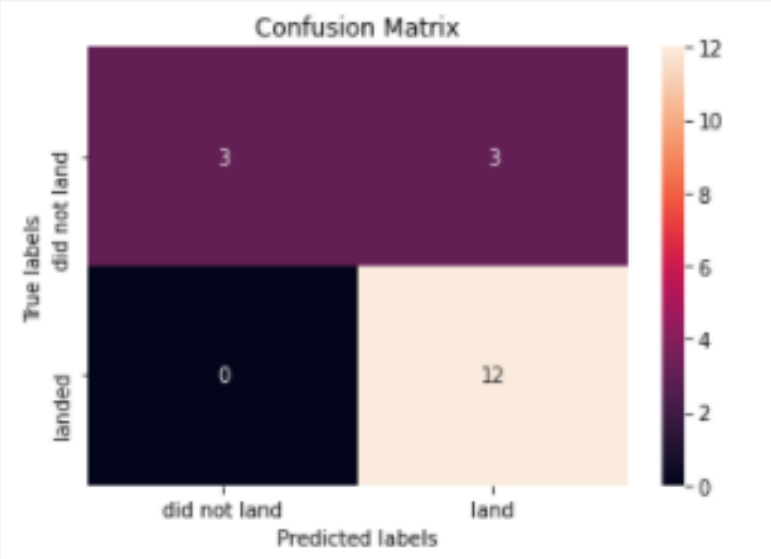
Decision Tree

Counts of prediction:

True: 16

False Positive: 1

False Negative: 1



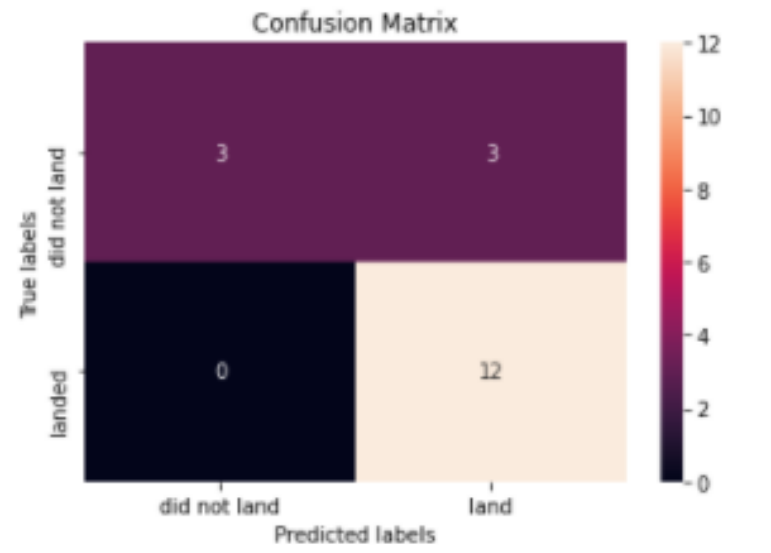
SVM

Counts of prediction:

True: 15

False Positive: 3

False Negative: 0



KNN

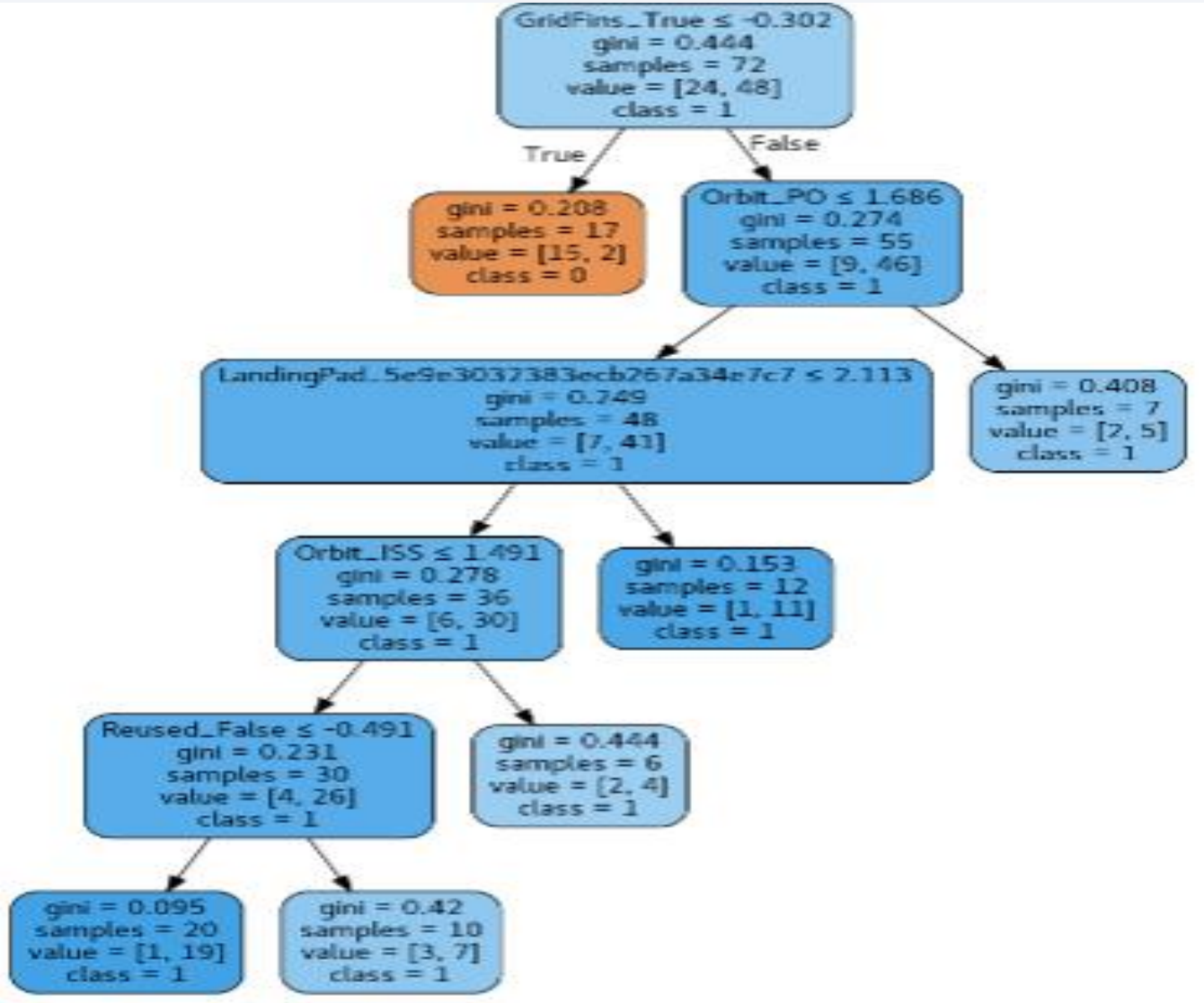
Counts of prediction:

True: 15

False Positive: 3

False Negative: 0

4.3.(iii). Optimized Decision Tree



- Optimized decision tree
- The flight without gridfins is more likely to fail
- After controlling all major factors to determine outcome, whether reusing first stage or not does not affect the outcome
- Reusing first stage is the good strategy to take for reducing cost of launch

5. Conclusions

Q1. What can determine the rocket launch outcomes; success or failure?

- Launch site, payload mass, orbit,
- The flight without gridfins is more likely to fail

Q2. After controlling all other factors determining outcome, does the reuse of the first stage affect the outcome of the rocket launch?

- No, after controlling all major factors to determine outcome, whether reusing first stage or not does not affect the landing outcome

Q3. Is reusing first stage worth?

- Reusing first stage is the good strategy to take for reducing cost of rocket launch in terms of statistical analysis of launch outcome

Whether SpaceY should reuse the first stage or not depends on:

- The actual cost of reusing the first stage; collection and maintenance cost vs producing the new
- Durability; how many times we can reuse without affecting launch outcome etc.

6. Appendix

- SpaceX API data source: https://api.spacexdata.com/v4/launches/past
- SpaceX web scraping data source: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- GitHub repository: https://github.com/mayuko-kondo/spacex_data_science
- Dashboard URL: https://mayukokondom-8050.theiadocker-5-labs-prod-theiak8s-4-tor01.proxy.cognitiveclass.ai/
- This is the final assignment of Applied Data Science Capstone course at Coursera (https://www.coursera.org/learn/applied-data-science-capstone?)

Thank you!